



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Identification and Characterisation of Putative Enhancer Elements in Mouse Embryonic Stem Cells

Citation for published version:

Mantsoki, A, Parussel, K & Joshi, A 2021, 'Identification and Characterisation of Putative Enhancer Elements in Mouse Embryonic Stem Cells', *Bioinformatics and biology insights*, vol. 15.
<https://doi.org/10.1177/1177932220974623>

Digital Object Identifier (DOI):

[10.1177/1177932220974623](https://doi.org/10.1177/1177932220974623)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Bioinformatics and biology insights

Publisher Rights Statement:

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Identification and Characterisation of Putative Enhancer Elements in Mouse Embryonic Stem Cells

Anna Mantsoki¹, Karla Parussel¹ and Anagha Joshi²

¹Division of Developmental Biology, The Roslin Institute, The University of Edinburgh, Midlothian, UK. ²Computational Biology Unit, Department of Clinical Science, University of Bergen, Bergen, Norway.

ABSTRACT: Enhancer elements control mammalian transcription largely in a cell-type-specific manner. The genome-wide identification of enhancer elements and their activity status in a cellular context is therefore fundamental to understanding cell identity and function. We determined enhancer activity in mouse embryonic stem (ES) cells using chromatin modifications and characterised their global properties. Specifically, we first grouped enhancers into 5 groups using multiple H3K4me1, H3K27ac, and H3K27me3 modification data sets. Active enhancers (simultaneous presence of H3K4me1 and H3K27ac) were enriched for binding of pluripotency factors and were found near pluripotency-related genes. Although both H3K4me1-only and active enhancers were enriched for super-enhancers and a TATA box like motif, active enhancers were preferentially bound by RNA polII (s2) and were enriched for bidirectional transcription, while H3K4me1-only enhancers were enriched for RNA polII (8WG16) suggesting they were likely poised. Bivalent enhancers (simultaneous presence of H3K4me1 and H3K27me3) were preferentially in the vicinity of bivalent genes. They were enriched for binding of components of polycomb complex as well as Tcf3 and Oct4. Moreover, a 'CTTTCTC' de-novo motif was enriched at bivalent enhancers, previously identified at bivalent promoters in ES cells. Taken together, 3 histone modifications successfully demarcated active, bivalent, and poised enhancers with distinct sequence and binding features.

KEYWORDS: epigenetics, enhancers, gene expression

RECEIVED: October 4, 2019. **ACCEPTED:** October 26, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by grants from the Biotechnology and Biological Sciences Research Council (BBSRC) (grant nos. BB/P013732/1 and BB/M020053/1). The Trond Mohn Foundation (grant no. BFS2017TMT01) currently supports A.J. Funding for open-access charge: RCUK funding for the University of Edinburgh. A.J. is supported by the Bergen Research Foundation Grant no. BFS2017TMT01.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Anagha Joshi, Computational Biology Unit, Department of Clinical Science, University of Bergen, 5th Floor, Thormøhlensgt 55, N-5008, Bergen, Norway. Email: Anagha.Joshi@uib.no

Introduction

Embryonic stem (ES) cells have an indefinite proliferative life span and have a potential for both generating daughter cells with equivalent potential or cells ready for differentiation, that is, to establish unique gene expression profiles characteristic to different cell and tissue types during development. The delicate balance between self-renewal and the competency-to-differentiate, is achieved by epigenetic and transcription control. This process is tightly regulated by transcription factors and chromatin-associated proteins, which respond to signalling pathways which in turn responds to environmental cues.¹ Epigenetic and transcription factors bind genomic regions proximal (promoters) and distal (enhancers) to the gene transcription start site (TSS) to control gene regulation. Collaborative efforts such as the FANTOM² and Roadmap Epigenomics³ projects have now successfully built enhancer and promoter repertoires across hundreds of cell types with an estimated 1.4% of the human genome associated to putative promoters and about 13% to putative enhancers. The vast majority of disease susceptibility loci lie in non-coding genomic regions, particularly in enhancer regions^{4,5} and have been estimated to explain a greater proportion of the heritability for some disorders than variants in coding regions.^{6,7} Despite their known functional importance during development and disease, computational approaches towards

genome-wide identification of enhancer elements remains a major challenge to date. This is due to various factors, including enhancers are typically distal to the TSS with no fixed distance preference, can range from a few to few hundred bases length and tend to be cell-type specific.⁸ On the other hand, regulatory regions (enhancers and promoters) have sequence features enriched for motifs of known transcription factors and evolutionarily conserved facilitate computational prediction. Moreover, enhancers can be distinguished from promoters using the presence of certain regulatory proteins including p300 (an acetyltransferase), BRG1 (a chromatin remodelled), the presence of histone modifications (H3K4me1 and H3K27ac), and unstable nucleosomes.⁹ Interestingly, several experiments have recently shown that enhancer regions get transcribed to form short, often bidirectional transcripts, called enhancers RNAs or eRNAs.¹⁰ Enhancer RNA expression is correlated with the expression of target genes, many a times in a stimulus-dependent manner and therefore can be used to predict enhancers and their target genes.¹¹ Several functional mechanisms of gene regulation have been described in literature including enhancer-promoter looping, recruiting transcriptional machinery, and facilitating RNA polymerase pause-release.¹² It is important to note that, eRNAs might not always be a predictor of enhancer activity or localisation, and not all functional enhancers have to be transcribed in eRNAs.



Enhancer regions have been classified into different groups based on their length, motif content, or chromatin status.¹³ Super-enhancers were defined as large genomic regions (>3 kb) near highly expressed genes, differing from typical enhancers in size, transcription factor density and content, ability to activate transcription, and sensitivity to perturbation.¹⁴ But, a CRISPR/Cas9-mediated deletion several super-enhancers clusters and isolated enhancers in mouse ES cells demonstrated enhancers and super-enhancers have an equivalent regulatory role in ES cells¹⁵ and iPS cells.¹⁶ The chromatin state is thought to be more functionally relevant with positive feedback loop between permissive chromatin and transition during early embryonic development.¹⁷ It is important to note that though histone modifications are widely used to define chromatin functional state, they are only predictions and not a proven fact. Furthermore, the experimental validations have demonstrated that they contain large number of false positives.^{16,18} Enhancers are marked by putative active (H3K4me1) or putative repressive (H3K27me3) chromatin modifications according to the cellular context. H3K4me3 (putative active) and H3K27me3 (putative repressive) modifications were found co-localised at promoters and enriched in developmentally regulated genes in mouse ES cells.¹⁹ Bivalent promoters were subsequently identified in human ES cells and were similarly enriched for developmental regulators²⁰ and conserved their chromatin signature across species.²¹ The bivalent state is thought to assure silencing of key developmental controllers in ES cells while keeping them poised for activation during differentiation.²² Similar to promoters, putative enhancers marked with the bivalent enhancers are defined by the presence of the repressive mark H3K27me3 and active mark H3K4me1. Bivalent promoters identified in many mature lineages including T-cells,²³ neural progenitor cells (NPCs), and mouse embryonic fibroblast (MEFs).²⁴ Similarly, putative bivalent enhancers were identified in pro-B cells and adult liver,²⁵ fibroblast-derived adipocytes, and bone-marrow macrophages.²⁶ In summary, bivalent domains are present in enhancers and promoters and are thought to prime regulatory regions for their differentiation trajectory.²⁷ An integrative analysis of bivalent promoters in mouse ES cells demonstrated that most H3K27me3 promoters were bivalent.²⁸ Importantly, most bivalent promoters did not resolve into active or repressed chromatin in differentiated lineages, with many bivalent chromatin domains in fact remained bivalent across diverse cell types.²⁹

Similarly, to study the functional relevance of chromatin signature of enhancer elements, we performed systematic analysis of putative enhancers in mouse ES cells by data integration. Specifically, we first grouped putative enhancers into 5 groups using multiple H3K4me1, H3K27ac, and H3K27me3 modification data sets. Putative active enhancers (simultaneous presence of H3K4me1 and H3K27ac) were enriched for binding of pluripotency factors and were found near pluripotency-related genes. Although both putative H3K4me1-only and

active enhancers were enriched for super-enhancers and a TATA box like motif, putative active enhancers were preferentially bound by RNA polII s2 form and were enriched for bidirectional transcription, while putative H3K4me1-only enhancers were enriched for RNA polII (8WG16) suggesting they were likely poised. Putative bivalent enhancers (simultaneous presence of H3K4me1 and H3K27me3) were preferentially in the vicinity of bivalent genes. They were enriched for binding of components of polycomb complex as well as Tcf3 and Oct4. Moreover, a 'CTTTCTC' de-novo motif was enriched at putative bivalent enhancers, previously identified at bivalent promoters in ES cells.

Material and Methods

Data collection and processing

We collected ChIP sequencing raw data for H3K4me1, H3K27ac, and H3K27me3 profiles with 4 samples of each type in mouse ES cells from GEO³⁰ and Roadmap Epigenomics.³¹ The accession numbers (containing the information about the ES cell lines and culture conditions) for raw data for each sample are provided in Supplemental Table S1. After mapping the reads to mm10 genome assembly, we called peaks in each sample using SICER.³² Input controls were not used, if they were not available for the samples. Specific parameters for running SICER for H3K4me3 and H3K27ac *window*=200 and *gap size*=200 and for H3K27me3, *window*=200 and *gap size*=2 × 300, since H3K27me3 covers wider chromatin domains. The rest of the parameters (same for H3K4me3, H3K27ac, and H3K27me3) were *effective genome fraction*=0.7, *redundancy threshold*=1, *fragment size*=150 and *E-value*=100. We filtered out GENCODE M2 TSSs, which include both protein-coding and non-coding genes, to select enhancer peaks (Supplemental Table S1). Promoter regions were defined as 1 kb region around GENCODE TSSs.

Putative enhancer classification

We created a genomic region by sample matrix by merging all peak files where '1' represented a peak present in a given genomic region in a given sample and '0' otherwise. Specifically, H3K4me1 peaks across diverse tissue and cell lines from ENCODE mouse data and 4 samples ES cell were merged using the following schema (Supplemental Table S2). The H3K4me1 peaks with an overlap of at least 25% with one of the peaks where the width of the merged peak was the union of 2 peaks. The merged list of H3K4me1 peaks was treated as 'putative enhancers' forming the rows of the matrix described above. The peaks for H3K4me1, H3K27ac, and H3K27me3 in ES cells overlapping at least 25% of the 'putative enhancer' peaks were considered overlapping and the H3K4me3 peak was defined as putative enhancer region. We then identified high confidence putative active enhancers by simultaneous presence of H3K4me1 and H3K27ac in 3 or more sample pairs

and absence of H3K27me3 in all 4 samples and high confidence putative bivalent enhancers by simultaneous presence of H3K4me1 and H3K27me3 in 3 or more sample pairs and absence of H3K27ac in all 4 samples. We also defined H3K4me1-only, H3K27ac-only, and H3K27me3-only peaks by the presence of each chromatin modification in 3 or more samples and absence of other modifications. We have provided the genomic co-ordinates for all predicted enhancers in each enhancer group as Supplementary Material 2.

Transcription and epigenetic control

To identify transcription and epigenetic factors enriched at putative bivalent, active, and H3K4me1-only enhancers, we used data from over 150 ChIP-seq experiments for transcription and epigenetic factors in mouse ES cells³³ and the significance of overlap was calculated using a hypergeometric test. Enriched regions or peaks from ChIP sequencing data for epigenetic and transcription factors in mouse ES cells were downloaded from CODEX database.³³ If the putative enhancer region from this study overlapped more than 50% with TF binding peaks from CODEX database, the peak was associated with the putative enhancer, that is, transcription factor was assumed to occupy that putative enhancer region. We collected ChIP sequencing data for 3 RNAPII modifications: RNAPIIS5P, RNAPIIS7P and 8WG16, PRC2 components: Suz12, Jarid2, and PRC1 subunits: Cbx7 and Ring1b in mouse ES cells and the enrichment of these factors at putative bivalent enhancers was calculated using BEDtools³⁴ and density plots were visualised using *R*. The TSS sites from FANTOM5 were downloaded from <http://fantom.gsc.riken.jp/5/data/>.

Enrichment analysis

The promoter regions in mouse ES cells grouped according to chromatin marks as bivalent, active (H3K4me3-only), H3K27me3-only, and latent (absence of both marks) were obtained from Mantsoki et al.²⁸ The data for ES cell putative enhancers in 2i culture³⁵ and ES cell putative super-enhancers¹⁴ was obtained from respective publications. The sequence motif enrichment analysis, the genomic location grouping (intergenic, intron, exon, TTS 3' or 5') as well as distance to the nearest TSS was calculated using HOMER.³⁶ The de-novo motif analysis was performed using HOMER with default parameters. Mammalian conservation tracks were downloaded from UCSC genome browser and conservation status was assigned if more than 75% of region overlapped with mammalian conserved regions. We used a stringent threshold to select the enriched motifs (motif must be present in at least 10% of sequences and with a *P* value < 1e-10). Gene ontology functional analyses for the putative bivalent, active, and H3K4me1-only enhancers were done using DAVID.³⁷ The TSSs predicted by the CAGE data as putative enhancers were obtained from the FANTOM consortium.² The enrichments of overlaps were

calculated using a hypergeometric test and *P* values were corrected using Bonferroni correction. The analysis was performed using *R* and shell scripts. All analysis was performed in mm10 genome assembly. The data sets provided by authors in mm9 were converted to mm10 using the liftover tool.³⁸

Putative enhancers in other tissues and during reprogramming

The peaks for the chromatin marks (H3K4me1, H3K27me3, and H3K27ac) across mouse cell lines and tissues were downloaded from the mouse ENCODE resource <https://genome.ucsc.edu/encode/downloadsMouse.html> (please refer to Supplemental Table S3 for details). The peaks for the chromatin marks (H3K4me1, H3K27me3, and H3K27ac) during reprogramming by 4 transcription factors Oct4, Sox2, Klf4, and Myc (OSKM) towards pluripotency were obtained from GEO, GSE67520.³⁹

Results

Classification of putative enhancers in mouse ES cells by integrating ChIP-seq data

To identify and characterise putative enhancer elements in mouse ES cells, we collected ChIP sequencing data sets for H3K4me1, H3K27ac, and H3K27me3 modifications in mouse ES cells from independent studies (4 samples for each modification, Supplemental Table S1). Putative enhancer peaks in each sample were determined by removing the peaks identified using SICER³² overlapping with GENCODE⁴⁰ promoters. This resulted into about 120 thousand nonpromoter genomic regions with H3K4me1 modifications in at least 1 of 4 samples (Figure 1A). About half of these regions were occupied by H3K27ac in at least 1 of the 4 samples and about 20% of these regions were occupied by H3K27me3 in at least 1 of the 4 samples. Importantly, H3K27ac and H3K27me3 tended to be mutually exclusive. For example, of more than 9000 peaks with H3K4me1 and H3K27ac modifications in all 8 samples, more than 8000 had no H3K27me3 in any of the samples. Nevertheless, over 200 peaks were marked with both H3K27ac and H3K27me3 in at least 3 samples. We classified 21 725 peaks with simultaneous presence of H3K4me1 and H3K27ac in 3 or more sample pairs and absence of H3K27me3 in all 4 samples as high confidence putative active enhancers and 2935 peaks with simultaneous presence of H3K4me1 and H3K27me3 in 3 or more sample pairs and absence of H3K27ac in all 4 samples as high confidence putative bivalent peaks. A total of 16 406 peaks present in 3 or more H3K4me1 samples and not in any of H3K27ac or H3K27me3 samples were called putative H3K4me1-only enhancers. Similarly, 832 putative H3K27ac-only enhancers and 409 putative H3K27me3-only enhancers were defined (Figure 1B).

We mapped regulatory regions to the TSS of the nearest gene resulting into on average 2.5 regulatory regions mapped to

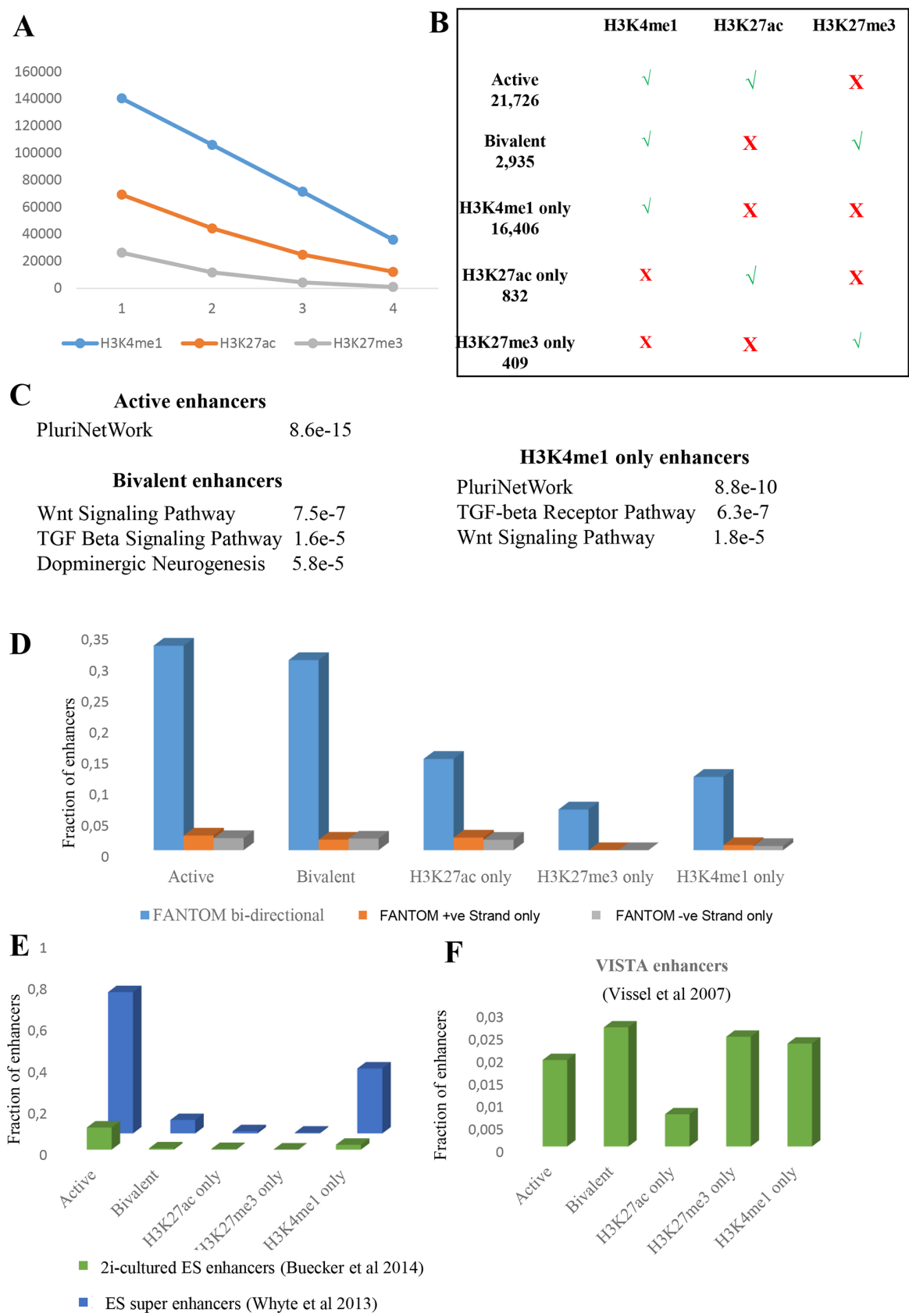


Figure 1. (A) The number of peaks in N (x axis) or more samples for each of the modification (H3K4me1 – red, H3K27ac – green, and H3K27me3–blue). (B) summary of enhancer groups based on the presence of histone modifications. (C) Gene ontology enrichment categories for the genes neighbouring bivalent, active and H3K4me1-only enhancers. (D) The fraction of transcription start sites in mouse ES cells identified by the FANTOM consortium overlapping with enhancer groups. (E) The fraction of enhancers in 2i- cultured ES cells and super-enhancers in ES cells overlapping with different enhancer groups. (F) The fraction of predicted by VISTA enhancer browser overlapping with different enhancer groups.

a gene. We then calculated functional enrichment and pathway enrichments for each gene list (Figure 1C). Gene neighbouring 'putative active enhancers' were enriched for pluripotency network genes (P value $< 8.6\text{e-}15$) and genes neighbouring 'putative bivalent enhancers' were enriched for Wnt Signalling Pathway (P value $< 7.5\text{e-}7$) and TGF Beta Signalling Pathway (P value $< 1.6\text{e-}5$). Interestingly, genes neighbouring 'putative H3K4me1-only enhancers' were enriched for both pluripotency network genes (P value $< 8.8\text{e-}10$) and Wnt Signalling Pathway (P value $< 6.3\text{e-}7$) and TGF Beta Signalling Pathway (P value $< 1.8\text{e-}5$).

Andersson et al⁴¹ used bidirectional expression from CAGE data to determine putative enhancer regions across multiple human and mouse cell types. We collected 44 507 bidirectional TSSs as well as 51 807 (positive-strand) and 46 434 (negative-strand) unidirectional TSSs in ES cells from FANTOM5 resource.² We noted that significantly higher fraction of regulatory regions in all groups overlapped with bidirectional TSSs compared to unidirectional TSSs (Figure 1D). We noted that about a third of both putative active and bivalent enhancers showed bidirectional expression or in other words, the majority of the regulatory regions defined as active or bivalent using chromatin modifications were not identified using bidirectional transcription signature.

Putative enhancers defined by published ChIP sequencing data of ES cells cultured in 2i medium condition³⁵ showed a very small overlap with the enhancers in all 3 groups (Figure 1E). In contrast, putative 'super-enhancers' in ES cells¹⁴ were highly enriched in both active and H3K4me1-only marked enhancers (Figure 1E). We collected enhancers predicted across diverse tissues and cell types in mouse from VISTA enhancer database⁴² and noted that they showed a small overlap with all putative enhancer groups (Figure 1F). Notably, putative H3K27me3-only enhancers showed similar overlap as putative active and bivalent enhancers (Figure 1F). We further confirmed that predicted enhancers across tissues in fact had a small overlap with all possible putative enhancers regions in ES cells (over 140 K), suggesting most tissue-specific enhancers are likely latent (with no histone modification) in ES cells.

Putative bivalent enhancers are enriched near bivalent gene promoters

We further analysed whether specific genomic regions were preferred by putative enhancers. Most putative enhancers were present in either intergenic or intronic regions for all putative enhancer types (Figure 2A). Given that mammalian genome has far more intergenic than intronic sequence, introns were enriched for enhancer elements of all types. Specifically, putative active and H3K4me1-only enhancers were highly enriched for intronic regions (P value $< 1\text{e-}50$) while putative bivalent enhancers were enriched to a lesser extent (P value $< 1\text{e-}3$).

We downloaded genomic regions conserved across mammals from UCSC genome browser and noted that more than half of all putative enhancers were conserved across mammals.

This fraction did not significantly change for any of the enhancer categories (Figure 2B).

To evaluate whether different putative enhancers showed a bias for specific genomic distance with respect to the nearest TSS. We calculated distance to the nearest TSS from each putative enhancer and noted that most enhancers were promoter distal with over half more than 50 kb away from the nearest TSS (Figure 2C). About 10% of H3K4me1-only enhancers formed an exception being within 5 kb of the nearest TSS.

We obtained the gene list likely to be regulated by putative enhancers in each category by mapping the putative enhancers to the TSS of the nearest gene (Supplemental Table S5). We have previously identified high-confidence bivalent, active (H3K4me3 only), H3K4me3-only and latent (no mark) promoters in mouse ES cells.²⁸ To test whether specific enhancer types are enriched near specific promoter types, we calculated the gene overlap between enhancer and promoter categories. Putative bivalent enhancers were highly enriched near bivalent promoters (Figure 2D). This postulated that bivalent promoters and enhancers might be part of a wider bivalent chromatin structure.

Specific epigenetic and transcription factors are enriched at putative active and bivalent enhancers

Voigt et al⁴³ proposed that bivalency in ES cells was maintained due to occupancy of fewer transcription factors, while at active promoters, the H3K27me3 modification is averted by a high density of active transcription factors. We collected ChIP sequencing data for more than 150 factors in mouse ES cells³³ and noted that bivalent promoters were indeed occupied by fewer transcription factors compared to the H3K4me3-only or active promoters (Figure 3A, right). Interestingly though, we noted that more factors were bound at putative bivalent enhancers compared to putative active enhancers in mouse ES cells (Figure 3A, left). Both bivalent and active promoters were occupied by many more factors than putative bivalent and active enhancers.

Bivalent promoters belong to 2 distinct groups in mouse ES cells. The first class consists of domains where only PRC2 exists ('PRC2-only') and the second one, called PRC1-positive, where PRC2 domains are also occupied by PRC1. PRC1-positive, in contrast to PRC2-only, were more expansive bivalent regions, highly conserved, and consisted numerous developmental promoters.⁴⁴ We collected ChIP sequencing data for epigenetic and RNAPII modifications known to distinguish bivalent promoter subtypes.²⁸ Specifically, we grouped 2 RNAPII modifications: RNAPIIS5P and RNAPIIS7P, 8WG16 (an antibody recognising nonphosphorylated C-terminal domain), PRC2 components: Suz12, Jarid2, and PRC1 subunits: Cbx7 and Ring1b density at putative bivalent enhancers in mouse ES cells (Supplemental Figure S1). We noted that similar to promoters, a fraction of putative bivalent enhancers indeed showed a strong occupancy of polycomb and RNAPII. Never the less, these enhancers formed about 10% of

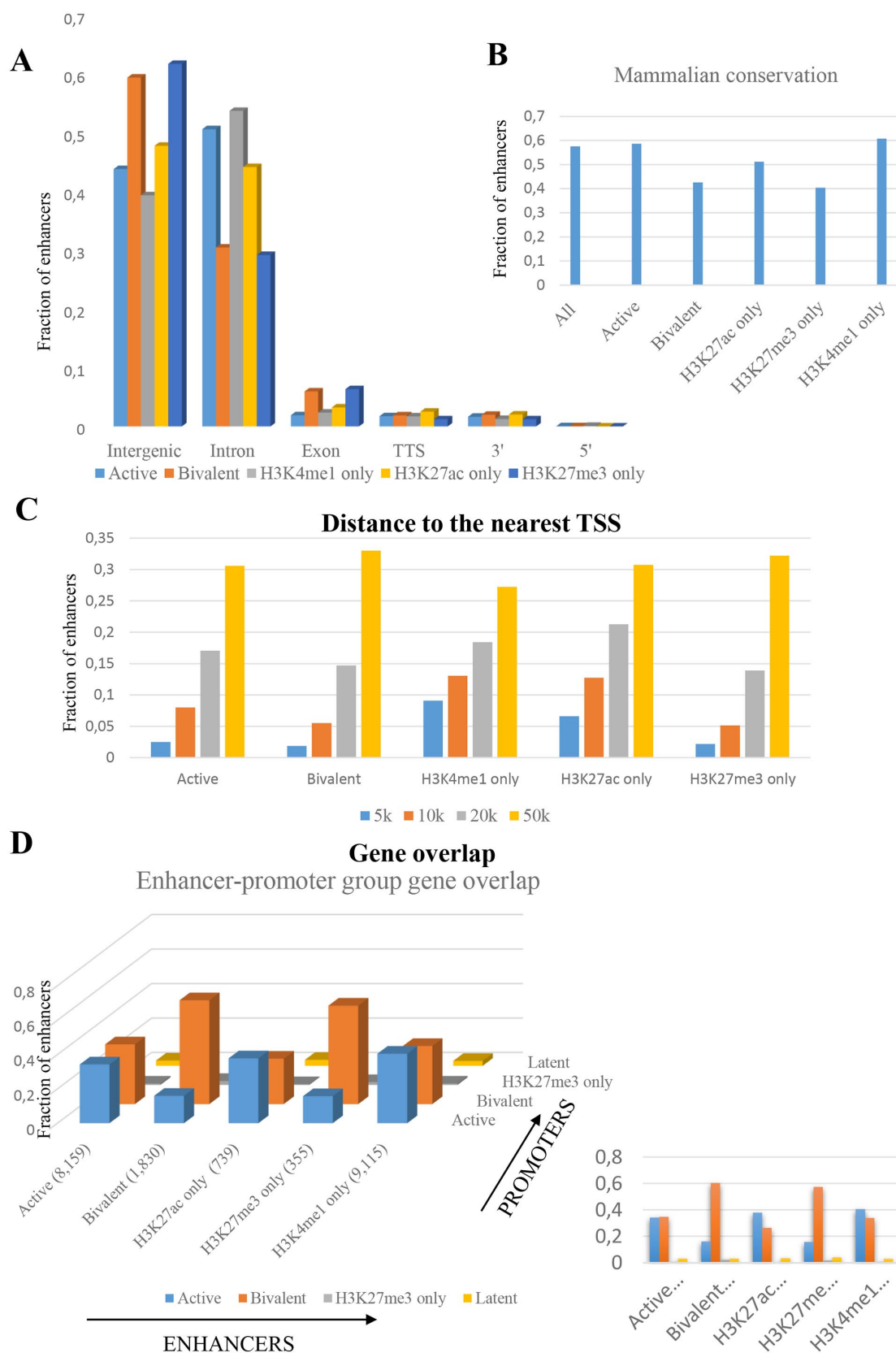


Figure 2. (A) The fraction each enhancer group in a specific genomic location (intergenic, intron, exon, 3' and 5') colour coded according to the enhancer group. (B) The fraction of each enhancer group overlapping with the mammalian conservation. (C) The fraction each enhancer group within a specific genomic distance window from the nearest TSS colour coded according to the distance window. (D) The overlap between the genes neighbouring the different promoter and enhancer groups representing bivalent enhancers and promoters are present near similar genes.

putative bivalent enhancers (Supplemental Figure S1), unlike promoters which formed about half of bivalent promoters.

We further calculated enrichment for binding of individual factors at each putative enhancer group (Supplemental Table S6). Active enhancers were significantly enriched for p300 and core pluripotency factors including Oct4, Sox2, Nanog, Smad1, and Stat3. Importantly, these factors were not enriched at putative H3K4me1-only enhancers (Figure 3B). Putative H3K4me1-only enhancers were specifically enriched for Rbbp5 and Dpy30, the common regulatory components of MLL complex. Putative bivalent enhancers were enriched for polycomb components Suz12 and Ezh2. Interestingly, they were also enriched for 2 pluripotency factors Tcf3 and Oct4.

As the number of binding events vary greatly across factors affecting the enrichment, we calculated the fraction of peaks in each group occupied by a given factor. Ezh2, Kdm2b, Rnf2, and Klf2 were present at about a third of putative bivalent enhancers while Nanog, Pou5f1, Brd4, and Prdm14 were present at about a quarter of putative active enhancers (Figure 3C). P300, together with Utx and Mll4, facilitates conversion of inactive enhancers to active enhancers in ESCs.⁴⁵ Interestingly, though p300 was enriched at putative active enhancers, it was present at the minority (less than 25%) of putative active enhancers. This can likely be due to the number of peaks called in p300 ChIP-seq sample (9429 peaks) compared with more than 21 000 predicted active enhancers.

Sequence motif enrichment at putative enhancer groups

In the previous section, we noted that putative active and bivalent enhancers were enriched for different epigenetic and transcription controller. To investigate whether the binding enrichment of these factors is also reflected in sequence motif enrichments at putative active and bivalent enhancers, we calculated known and de-novo motif enrichment using HOMER.³⁶ We selected motifs enriched with a *P* value less than $1e-10$ and present in at least 10% of enhancers. Both putative active and H3K4me1-only enhancers were enriched for KLF, FLI, and ETS motifs. Importantly, the sequence motif for pluripotency factors enriched at putative active enhancers was not highly enriched. This suggests that KLF, FLI, and ETS motifs might facilitate binding of pluripotency factors to the enhancers. Indeed, simultaneous depletion of Klf2, Klf4, and Klf5 lead to ES cell differentiation.⁴⁶ No known motif was found enriched at bivalent enhancers using stringent cut-offs defined above. The lack of enrichment of known pluripotency factor motifs at putative active enhancers and polycomb complexes components at bivalent enhancers might be due to wider peaks called in histone modification data.

The de-novo motif search identified 3 motifs enriched in bivalent enhancers, 3 in putative active enhancers and 2 in putative H3K4me1-only enhancers (Figure 4B). The first

de-novo motif identified at putative bivalent enhancers was previously identified enriched at bivalent promoters in ES cells.²⁸ The second de-novo motif was similar to Myf6 from JASPAR database,⁴⁷ while the third was present in about half of bivalent enhancers and was similar to Egr1 from JASPAR database.⁴⁷ Of the 3 motifs in putative active enhancers, the first was similar to a POLII TATA box, the second similar to KLF motif while the third similar to MA0006.1, a bHLH motif from JASPAR database.⁴⁷ Supporting known motif enrichments, the de-novo motif similar to TATA box was present in active and H3K4me1-only enhancers. Putative H3K4me1-only specific de-novo motif was similar to Zbtb3 motif from JASPAR database.⁴⁷ As POLII TATA box like motif was enriched in putative active and H3K4me1-only enhancers, we calculated the overlap between enhancers and 3 forms of RNA POLII binding (s2, s5, and 8GW16) using ChIP sequencing data.⁴⁸ Putative H3K4me1-only enhancers showed a much higher overlap with 8WG16 antibody, recognising nonphosphorylated Pol II while putative active enhancers showed a higher overlap for s2 antibody, specific for active, elongating Pol II (Supplemental Table S4).

ES putative enhancer status in mouse tissues and during cellular reprogramming

To study the dynamics of the 5 groups of putative enhancers in ES cells through differentiation, we collected H3K4me1, H3K27ac, H3K27me3 data for 10 tissues and cell types from the mouse ENCODE resource. Similar to the analysis of ES cells data, we classified the putative enhancers in each cell type in the same 5 groups, namely active, bivalent, H3K4me1-only, H3K27ac-only, and H3K27me3-only. We then calculated the overlap of each groups of putative enhancers in ES cells with all 5 enhancers groups in each of the 10 tissue and cell types individually (Figure 5, left). Putative active and H3K4me1-only enhancers in ES cells were mostly active or H3K4me1-only in other tissue of cell types (Figure 5A, C left). Putative bivalent enhancers in ES cells, though were predominantly H3K4me1-only in other tissue and cell types, about 10% retained bivalency across other tissue and cell types (Figure 5B, left). Interestingly, about 20% of putative bivalent enhancers in ES cells were H3K27ac-only in heart and liver (Figure 5B, left).

To explore how the above-noted patterns might establish during differentiation, we collected time-series data during cellular reprogramming by induced expression of 4 transcription factors Oct4, Sox2, Klf4, and Myc (OSKM) pushing somatic cell reprogramming towards pluripotency.³⁹ Similar to the observation in other tissues and cell types, putative active and H3K4me1-only enhancers in ES cells were mostly active or H3K4me1-only during reprogramming. As expected putative active enhancers in ES cells showed a high overlap with putative active enhancers in iPSCs and putative H3K4me1-only enhancers also behaved similarly. Interestingly, putative

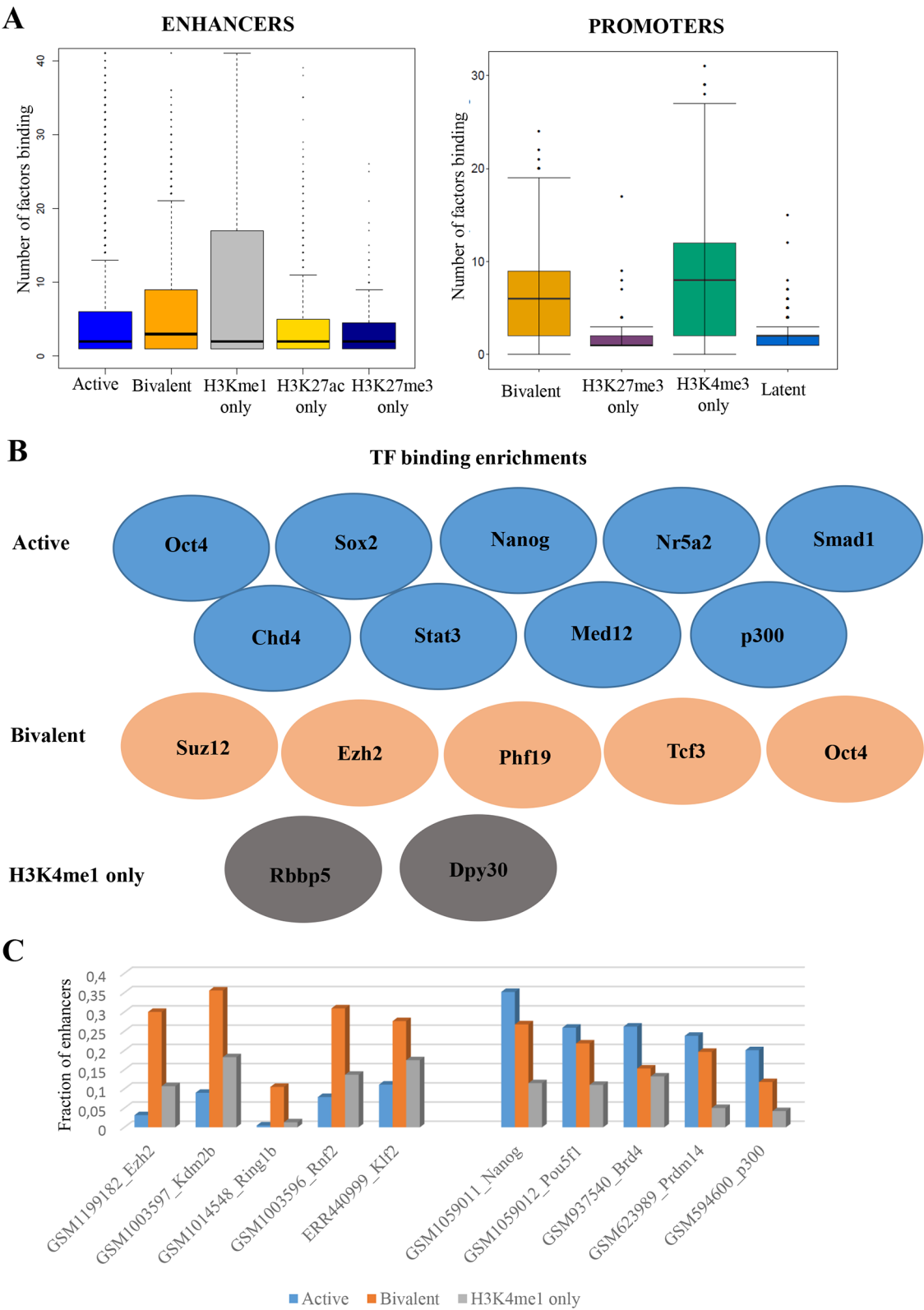








Figure 3. (A) The box plots of number of factors binding at different enhancer groups (left) and promoter groups (right) using CODEX database. (B) The epigenetic and transcription factors enriched at the active, bivalent, and H3K4me1-only enhancers. (C) The fraction of active, bivalent, and H3K4me1-only enhancers overlapping with peaks of different epigenetic and transcription factors. The factors at left are enriched in bivalent enhancers while the factors at right are enriched for the active enhancers.

A				
KNOWN MOTIF				
	Consensus sequence	Known TF	P value	% peaks
Active enhancers		Klf (Zf)	1e-35	14.3%
		Fli (Etv)	1e-15	11.1%
		Ets (Etv)	1e-12	14.6%
H3K4me1 only Enhancers		Ets (Etv)	1e-79	15.7%
		Klf (Zf)	1e-11	11.0%
		Fli (Etv)	1e-11	10.9%









B			
DE-NOVO MOTIF			
	Consensus sequence	P value	% peaks
Bivalent enhancers		1e-16	10.6
		1e-14	14.2
		1e-12	46.7
Active enhancers		1e-39	24.3
		1e-38	11.9
		1e-17	16.7
H3K4me1 only enhancers		1e-28	42.2
		1e-12	28.8

Figure 4. (A) Enriched known motifs at active and H3K4me1-only enhancers present in at least 10% of the enhancers identified by HOMER. (B) Enriched de-novo motifs at active and H3K4me1-only enhancers present in at least 10% of the enhancers identified by HOMER.

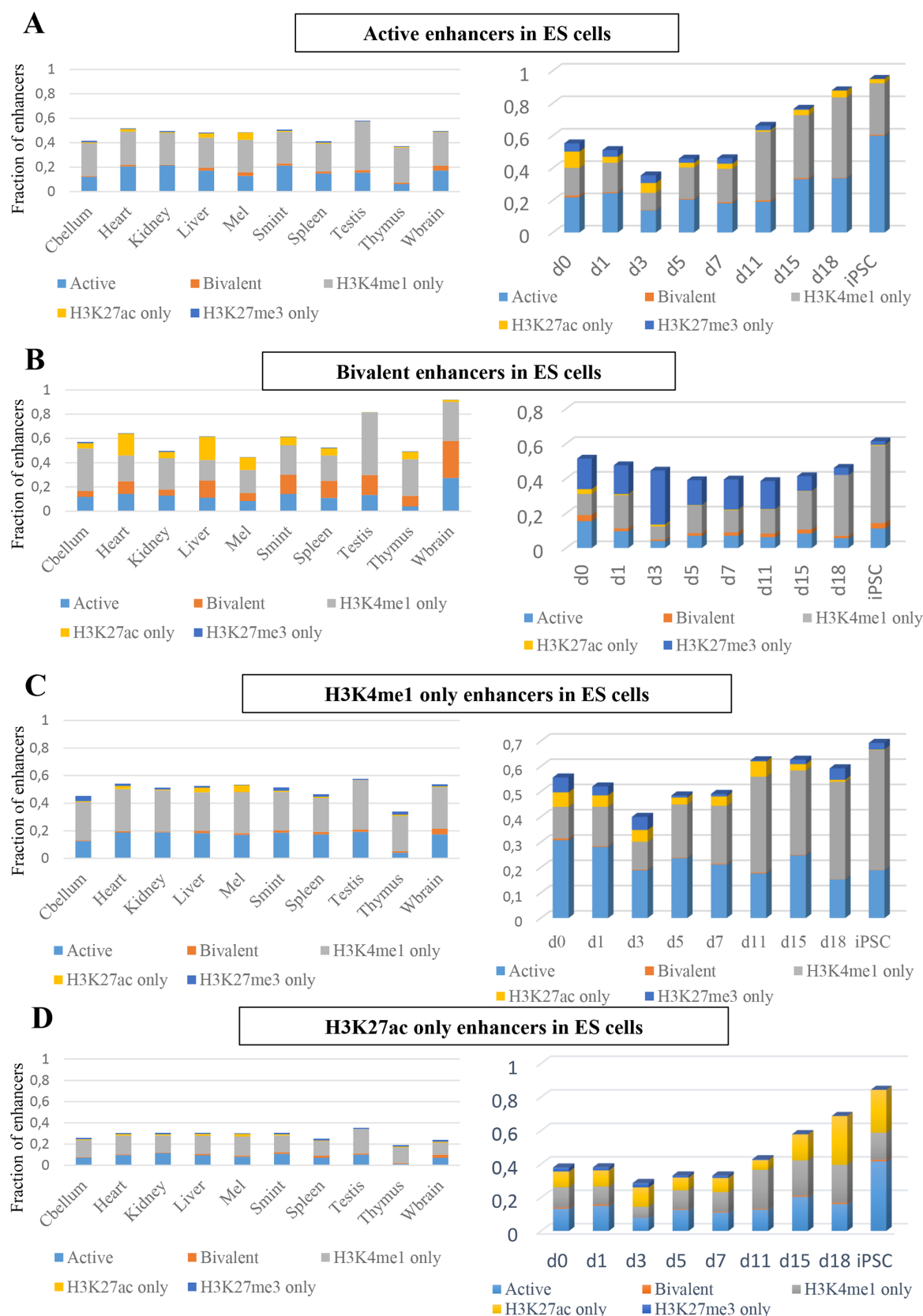


Figure 5. (A) The fraction of active enhancers overlapping with enhancer groups across different tissues from the mouse ENCODE (left) and with enhancers groups during cellular reprogramming (right) (Cbellum=cerebellum and Wbrain=whole brain). (B) The fraction of bivalent enhancers overlapping with enhancer groups across different tissues from the mouse ENCODE (left) and with enhancers groups during cellular reprogramming (right). (C) The fraction of H3K4me1-only enhancers overlapping with enhancer groups across different tissues from the mouse ENCODE (left) and with enhancers groups during cellular reprogramming (right). (D) The fraction of H3K27ac-only enhancers overlapping with enhancer groups across different tissues from the mouse ENCODE (left) and with enhancers groups during cellular reprogramming (right).

bivalent enhancers in ES cells lost H3K27me3 and gained H3K4me1 during reprogramming.

Discussion

Understanding enhancer function is fundamental to understanding cell-type-specific gene control. Here, we have defined putative enhancers based on the presence of H3K27ac, H3K27me3, and H3K4me1 modifications and characterised them in detail. We defined putative active enhancers by simultaneous presence of H3K4me1 and H3K27ac. We note that experimental validations of putative active enhancer predictions demonstrate that they contain high rate of false positives^{16,18} and have led to other definitions of active enhancers but there is no agreed consensus. Simultaneous presence of H3K4me1 and H3K27ac remains a common criterion used for identification putative active enhancers. We note that in this study used only histone modification data, as the aim was to identify high confidence (relatively wide) regulatory regions and not identification of precise regulatory site within them. In latter case, one would ideally integrate other data including p300 binding or nucleosome depletion measured via ATAC-Seq or DNase hypersensitivity assays. Furthermore, we mapped the regulatory regions to the nearest TSS. Although this is the most widely used approach to map putative targets of regulatory regions, it is important to note that the nearest neighbour gene is not always the enhancer target. There are many approaches to improve the mapping including chromatin capture assays to computational models.⁴⁹

Putative active enhancers showed a high overlap with FANTOM bidirectional putative enhancers, ES putative super-enhancers, and binding of many pluripotency factors. We note that the FANTOM bidirectional putative enhancers were identified across diverse tissue and cell types and therefore limiting the overlap. The properties of active enhancers were distinct from putative H3K4me1-only enhancers corroborating that the presence of H3K27ac is sufficient to distinguish putative active enhancers.²⁵ The higher overlap of putative H3K4me1-only enhancers with RNA POLII 8WG16 antibody, with no enrichment for bidirectional expression points out that they might be poised. Intriguingly, genes near H3K4me1-only enhancers were also enriched for pluripotency-related genes. Furthermore, genes neighbouring 'putative active enhancers' were enriched for pluripotency network genes (P value $< 8.6\text{e-}15$) and genes neighbouring 'putative bivalent enhancers' were enriched for Wnt Signalling Pathway (P value $< 7.5\text{e-}7$), and TGF Beta Signalling Pathway (P value $< 1.6\text{e-}5$). Interestingly, genes neighbouring 'putative H3K4me1-only enhancers' were enriched for both pluripotency network genes (P value $< 8.8\text{e-}10$) and Wnt Signalling Pathway (P value $< 6.3\text{e-}7$) and TGF Beta Signalling Pathway (P value $< 1.8\text{e-}5$). We noted that putative H3K4me1-only enhancers share the (nearest) genes with putative active and bivalent enhancers, resulting into enrichment for the signalling pathways. H3K4me1 modification has sharp ChIP-seq signal

compared to both H3K27ac and H3K27me3 modifications and therefore more consistent peak detection across data sets. One of the possibilities therefore is that at least some of the putative H3K4me1-only enhancers were in-fact putative active (missing H3K27ac) or bivalent enhancers (missing H3K27me3). It is also worth noting that bivalent enhancers had all 3 RNA Pol II peaks present (Supplemental Table S4).

We defined putative bivalent enhancers by simultaneous presence of H3K4me1 and H3K27me3 modifications. Putative bivalent enhancers were present near bivalent promoters and many maintained bivalency through differentiation. Other definitions of putative bivalent enhancers include simultaneous presence of H3K27ac and DNA methylation.⁵⁰ Bivalent regions selected by this definition show a high overlap with our 'putative active enhancers' as most H3K27ac marked enhancers also carry H3K4me1 modification. Finally, we also identified a 'CTTTCTC' de-novo motif enriched at bivalent enhancers previously identified at bivalent promoters.

Acknowledgements

The authors thank Guillaume Devailly for useful discussions.

Author Contributions

AM and KP performed the analysis. AM and AJ interpreted the results and AJ wrote the manuscript.

Data Availability

The raw data sets used in this manuscript are available in the public domain with identification numbers: SRR317224, SRR414939, SRR900935, SRR900936, SRR317225, SRR414936, SRR900947, SRR900948, GSE31039, GSE38596, GSE47949 (rep 2), and GSE47949 (rep 3).

ORCID iD

Anagha Joshi  <https://orcid.org/0000-0003-1184-3305>

Supplementary Material

Supplementary material for this article is available online for this article.

REFERENCES

1. Ng H-H, Surani MA. The transcriptional and signalling networks of pluripotency. *Nat Cell Biol.* 2011;13:490–496.
2. Forrest ARR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. *Nature.* 2014;507:462–470.
3. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007;447:799–816.
4. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518:317–330.
5. Perenthaler E, Yousefi S, Niggel E, Barakat TS. Beyond the exome: the non-coding genome and enhancers in neurodevelopmental disorders and malformations of cortical development. *Front Cell Neurosci.* 2019;13:352.
6. Corradin O, Scacheri PC. Enhancer variants: evaluating functions in common disease. *Genome Med.* 2014;6:85.
7. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 2012;337:1190–1195.
8. Devailly G, Joshi A. Insights into mammalian transcription control by systematic analysis of ChIP sequencing data. *BMC Bioinformatics.* 2018;19:409.
9. Kim SG, Harwani M, Grama A, Chaterji S. EP-DNN: a deep neural network-based global enhancer prediction algorithm. *Sci Rep.* 2016;6:38433.

10. Natoli G, Andrau JC. Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet.* 2012;46:1-19.
11. Arner E, Daub CO, Vitting-Seerup K, et al. Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science.* 2015;347:1010-1014.
12. Hou TY, Kraus WL. Spirits in the material world: enhancer RNAs in transcriptional regulation [published online ahead of print September 1, 2020]. *Trends Biochem Sci.* doi:10.1016/j.tibs.2020.08.007.
13. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012;9:215-216.
14. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell.* 2013;153:307-319.
15. Moorthy SD, Davidson S, Shchuka VM, et al. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* 2017;27:246-258.
16. Barakat TS, Halbritter F, Zhang M, et al. Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell.* 2018;23:276.e8-288.e8.
17. Bulut-Karslioglu A, Macrae TA, Oses-Prieto JA, et al. The transcriptionally permissive chromatin state of embryonic stem cells is acutely tuned to translational output. *Cell Stem Cell.* 2018;22:369.e8-383.e8.
18. Kwasniewski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* 2014;24:1595-1602.
19. Bernstein BE, Mikkelsen TS, Xie X, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell.* 2006;125:315-326.
20. Pan G, Tian S, Nie J, et al. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell.* 2007;1:299-312.
21. Zhao XD, Han X, Chew JL, et al. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell.* 2007;1:286-298.
22. Azuara V, Perry P, Sauer S, et al. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol.* 2006;8:532-538.
23. Roh T-Y, Cuddapah S, Cui K, Zhao K. The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci USA.* 2006;103:15782-15787.
24. Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature.* 2007;448:553-560.
25. Creighton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci.* 2010;107:21931-21936.
26. Zentner GE, Tesar PJ, Schacher PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* 2011;21:1273-1283.
27. Mohn F, Weber M, Rebhan M, et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell.* 2008;30:755-766.
28. Mantsoki A, Devailly G, Joshi A. CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells. *Sci Rep.* 2015;5: 16791.
29. Mantsoki A, Devailly G, Joshi A. Dynamics of promoter bivalency and RNAP II pausing in mouse stem and differentiated cells. *BMC Dev Biol.* 2018;18:2.
30. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res.* 2013;41:D991-D995.
31. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol.* 2010;28:1045-1048.
32. Xu S, Grullon S, Ge K, Peng W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol Biol Clifton NJ.* 2014;1150:97-111.
33. Sanchez-Castillo M, Ruau D, Wilkinson AC, et al. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* 2015;43:D1117-D1123.
34. Quinlan AR, Hall IM. {BEDTools:} a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841-842.
35. Buecker C, Srinivasan R, Wu Z, et al. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell.* 2014;14:838-853.
36. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38:576-589.
37. Dennis G, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 2003;4: P3.
38. Karolchik D, Barber GP, Casper J, et al. The UCSC genome browser database: 2014 update. *Nucleic Acids Res.* 2014;42:D764-D770.
39. Chen J, Chen X, Li M, et al. Hierarchical Oct4 binding in concert with primed epigenetic rearrangements during somatic cell reprogramming. *Cell Rep.* 2016;14:1540-1554.
40. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012;22:1760-1774.
41. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature.* 2014;507:455-461.
42. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* 2007;35: D88-92.
43. Voigt P, Tee W-W, Reinberg D. A double take on bivalent promoters. *Genes Dev.* 2013;27:1318-1338.
44. Ku M, Koche RP, Rheinbay E, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* 2008;4:e1000242.
45. Wang S-P, Tang Z, Chen C-W, et al. A UTX-MLL4-p300 transcriptional regulatory network coordinately shapes active enhancer landscapes for eliciting transcription. *Mol Cell.* 2017;67:308.e6-321.e6.
46. Jiang J, Chan Y-S, Loh Y-H, et al. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol.* 2008;10:353-360.
47. Khan A, Fornes O, Stigliani A, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 2018;46:D260-D266.
48. Das PP, Shao Z, Beyaz S, et al. Distinct and combinatorial functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in mouse embryonic stem cell identity. *Mol Cell.* 2014;53:32-48.
49. Fulco CP, Nasser J, Jones TR, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet.* 2019;51:1664-1669.
50. Charlet J, Duymich CE, Lay FD, et al. Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol Cell.* 2016;62:422-431.